Bounding probabilities in a Markovian model of character evolution

by

Christopher Tuffley
Supervised by Dr Michael Steel
Department of Mathematics and Statistics,
University of Canterbury, Christchurch, New Zealand

No. 132 November, 1995

Abstract: We investigate a probability-based model of character evolution. Using Erdös-Székely path systems, a sharp upper bound for the probability of a given character evolving under this model is obtained in terms of its parsimony length. In the case of binary characters the conditions under which this bound is realised are determined.

Contents

1 Introduction

This report summarises the work I did in my honours project in the field of phylogenetic analysis, supervised by Dr Steel. Phylogenetic analysis is an applied branch of mathematics used in the study of classes of objects that have or are assumed to have evolved on some sort of tree-like structure, such as languages or species of animals. A familiar example of such a structure is a family tree, though strictly speaking these do not form trees unless we consider only the male or female lines.

One aim of phylogenetic analysis is to use information from the "leaves" to determine the structure of the underlying tree. While a genealogist may use records of births and deaths to tackle the equivalent problem of reconstructing a family tree, in other cases of interest such records may not exist, and we may be

forced to use characteristics such as eye colour or presence of a particular gene to guess at the relationships between our objects of study. Information such as the eye colour of every living individual we wish to include on our tree is called a *character*, and is the data used in phylogenetic analysis to reconstruct trees. For example, by modeling the way in which a character "evolves" on a tree, we could assign each character a probability of evolving on a given tree, and then choose the tree on which our character is most likely to evolve.

In this project I studied such a model of character evolution. The starting points of this investigation were two conjectures of Dr Steel's:

1. An upper bound on the probability of a character evolving in terms of its "length" already established for characters that are assumed to take one of only two possible values at each site (two colour characters) would generalise in a natural way to characters that are assumed to take one of r possible values at each site (r colour characters).

2. The parameters maximising the probability of a two colour character evolving are related in a simple way to certain "extensions" of the character.

I established both of these conjectures, which appear in this report as Theorems 3 and 6. In the process, I also established a number of other results of theoretical interest. Work on both problems illustrated the usefulness of Menger's theorem and a generalisation due to Erdős and Székely, which relate the length of a character to certain systems of paths within the tree.

Section 2 of this report formalises the concepts outlined above and introduces the model of character evolution and some of the theory we will be using. Conjectures 1 and 2 above are dealt with in sections 3 and 4 respectively, and in sections 5 and 6 applications of the results of this project are considered and some open questions are posed.

2 Preliminaries

2.1 Definitions

Definitions 1 (Phylogenetic trees, characters) *A* phylogenetic tree *is a tree $T = (V(T), E(T))$ having no vertices of degree two and such that each* leaf *(degree one vertex) is given a unique label from $\{1, \ldots, n\}$, where n is the number of leaves of T. We say that T is a tree on n leaves, and write $[n]$ for $\{1, \ldots, n\}$. Where convenient, we identify each leaf with its label. If every internal (non-leaf) vertex of T has degree three, we say that T is* binary. *In the case of rooted trees, we allow the root to have degree two.*

A function $\chi : [n] \mapsto C$, where C is a set of r colours, *is an* (r-colour) character. *When $r = 2$, χ is said to be* binary. *A function $\hat{\chi} : V(T) \mapsto C$ is*

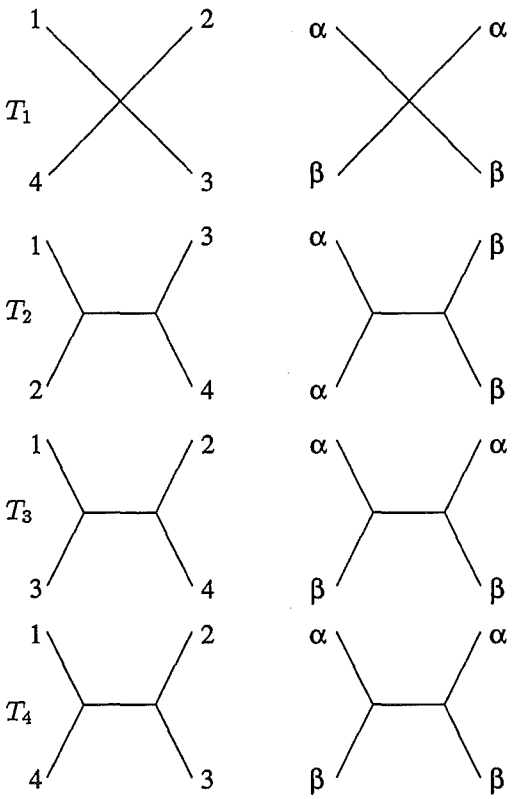


Figure 1: Some examples of phylogenetic trees.

called a colouration *of T; if $\hat{\chi}$ is such that $\hat{\chi}|_{[n]} = \chi$ (that is $\hat{\chi}$ agrees with χ on the leaves of T) then $\hat{\chi}$ is called an* extension *of χ (on T).*

Notation: The edge incident with the vertices u and v will usually be denoted by $\{u, v\}$. However, where we consider this edge to be directed from u to v it will be denoted by the ordered pair (u, v).

Figure 1 shows the four (unrooted) phylogenetic trees on four leaves and the way in which the binary character

i	1	2	3	4
$\chi(i)$	α	α	β	β

(1)

appears on each of these trees. The trees T_2, T_3 and T_4 are binary. Note that these three trees are considered to be distinct even though they are isomorphic as graphs, as the graph isomorphism does not preserve the leaf labelling.

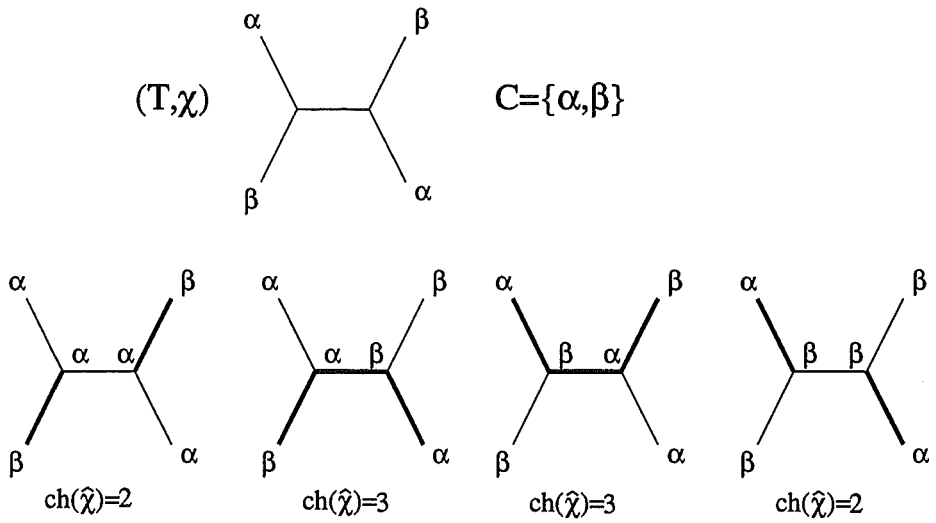


Figure 2: To find $\ell(\chi, T)$ for the tree and character shown with colour set $C = \{\alpha, \beta\}$, consider all possible colourations of the internal vertices. Since T has two internal vertices, there are $2^2 = 4$ such colourations. Bi-coloured edges are shown bolded. The minimum value of $ch(\hat{\chi})$ is two, so that $\ell(\chi, T) = 2$. There are two minimal extensions.

With each character χ and phylogenetic tree T on n leaves we may associate a non-negative integer (the "length" of χ on T) as follows.

Definitions 2 (length of χ on T, minimal extensions) *If $\hat{\chi} : V(T) \mapsto C$ then the* changing number *of $\hat{\chi}$, $ch(\hat{\chi})$, is the number of edges $e = \{u, v\}$ such that $\hat{\chi}(u) \neq \hat{\chi}(v)$. Such an edge is said to be* bi-coloured.

If $\chi : [n] \mapsto C$ then the length of χ on the phylogenetic tree T, $\ell(\chi, T)$, *is the minimum of $ch(\hat{\chi})$ over all extensions $\hat{\chi}$ of χ on T. An extension of minimal changing number is called a* minimal extension of χ *(on T).*

Biologically, we interpret each vertex of a phylogenetic tree as representing a species, with the edges denoting (immediate) ancestor-descendant relationships. The leaves represent extant species, the internal vertices ancestral species, and in rooted trees the root represents a common ancestral species from which all other species on the tree are descended. Since we are primarily interested in speciation events, where the tree "branches", we do not allow vertices of degree two except possibly at the root.

Characters are obtained by gathering data such as DNA sequence information from present day species. Each extension of a character is a way that it could have evolved on the tree, and the changing number of an extension is the number of changes or mutations it involves. The length of a character is

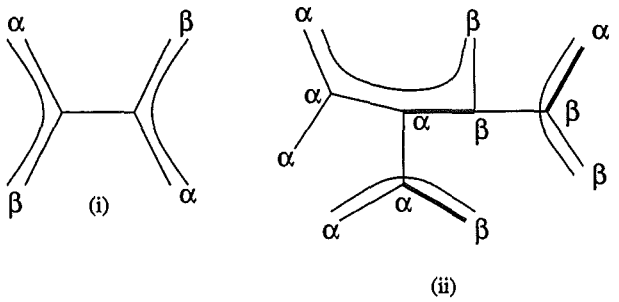


Figure 3: Some path systems illustrating Menger's theorem. (i) A maximal path system for the tree and character in figure 2. (ii) A set of three edge-disjoint paths joining differently coloured leaves, together with an extension of changing number three. By Menger's theorem the character shown has length three on the given tree.

therefore the minimum number of mutations required for it to evolve on the tree, and is used in methods such as maximum parsimony to estimate the true phylogeny of extant species.

2.2 Menger's theorem and Erdős-Székely path systems

In practical applications the length of a character on a given tree is found using Fitch's algorithm, which is an order n process for determining $\ell(\chi, T)$ and finding a minimal extension. However, for theoretical purposes $\ell(\chi, T)$ is usefully given by Menger's theorem and Erdős-Székely path systems, two results that will be of great importance to us in later sections. We state them here in the form in which we will be using them, rather than in their full generality.

Theorem 1 (Menger's theorem for trees, [1]) *If χ is a binary character then $\ell(\chi, T)$ equals the maximum number of edge-disjoint paths connecting differently coloured leaves of T.*

Although Menger's theorem applies only to binary characters, an extension to r-colour characters has been developed recently by Erdős and Székely [5].

Definitions 3 (Erdős-Székely path systems) *An* Erdős-Székely path system *for χ on T is a set $\mathcal{P}$ of directed paths in T satisfying the following conditions:*

1. *Each path joins leaves coloured differently by χ.*
2. *If two paths use the same edge of T, then*
   (a) *they use it in the same direction, and*

(b) they are directed towards leaves coloured differently by χ.

If $\mathcal{P}$ has the maximum cardinality of any Erdős-Székely path system for χ on T, then $\mathcal{P}$ is said to be optimal.

Notation: Following Erdős and Székely [5] we denote the starting vertex of a directed path P by $s(P)$ and the terminal vertex of P by $t(P)$.

Theorem 2 (Erdős and Székely [5]) *The size of an optimal Erdős-Székely path system for χ on T equals $\ell(\chi, T)$.*

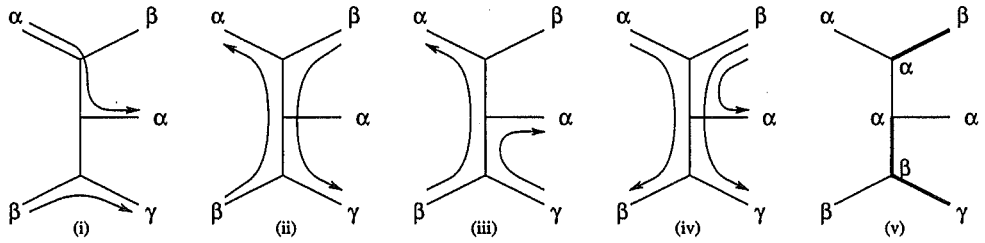


Figure 4: Some path systems and a colouration illustrating Erdős-Székely path systems and Theorem 2. (i)–(iv) show some path systems, of which only (iv) is an Erdős-Székely path system. (v) shows a minimal colouration.

Figure 4 illustrates Erdős-Székely path systems and Theorem 2. Only one of the path systems shown is an Erdős-Székely path system: (i) contains a path connecting leaves that are the same colour α, violating condition 1; in (ii) each path joins differently coloured leaves, but the same edge is used in opposite directions by two paths, breaking condition 2(a); while in (iii) condition 2(a) is satisfied but 2(b) is not as both paths share an edge and are directed towards identically coloured leaves. (iv) is an Erdős-Székely path system since all of the conditions are satisfied. The colouration in (v) has changing number three, and it follows from Theorem 2 that the path system in (iv) is optimal and that the character has length three on the given tree.

Note that although Theorem 2 includes the case $r = 2$, it does not reduce to Menger's theorem when χ is binary, as it allows the paths to intersect. Viewed in the light of Theorem 2, Menger's theorem guarantees us the existence of an edge-disjoint Erdős-Székely path system when $r = 2$, a fact we will make use of in section 4.

2.3 A Markovian model of character evolution

The model we will be considering is a generalisation to r colours of the Cavender-Farris (two colour case, [3, 6]) and Jukes-Cantor (four colour case, [10]) models.

Given a rooted phylogenetic tree T, a probability distribution π of colours at the root ρ and a mutation probability p_e on each edge e of T, the colour at the root "evolves" down the tree, assigning a colour to each vertex of T and generating a colouration $\hat{\chi}$ of T. We suppose that this evolution takes place such that

- there is a total order $\leq$ of the vertices, respecting ancestry (so $u < v$ if u is nearer the root than v is), such that

$$\mathbb{P}[\hat{\chi}(v) = \alpha | \bigwedge_{w<v} \hat{\chi}(w)] = \mathbb{P}[\hat{\chi}(v) = \alpha | \chi(w_0)] \quad \forall \alpha \in C, v \in V(T) \tag{2}$$

  where w_0 is the immediate ancestor of v

- the probability of a net change of colour occurring across an edge e is given by p_e, and if a net change occurs, each of the remaining $r-1$ colours is equally likely

- p_e satisfies $0 \leq p_e \leq (r-1)/r$.

The probability of generating a given colouration $\hat{\chi}$ will in general depend on T, π and the vector $p = (p_e)_{e \in E(T)}$ of probabilities and is given by

$$\mathbb{P}[\hat{\chi}|T, \pi, p] = \pi(\hat{\chi}(\rho)) \prod_{\substack{e=\{u,v\}: \\ \hat{\chi}(u)=\hat{\chi}(v)}} (1-p_e) \prod_{\substack{e=\{u,v\}: \\ \hat{\chi}(u)\neq\hat{\chi}(v)}} \frac{p_e}{r-1} . \tag{3}$$

The probability of generating a character χ under the model is found by summing (3) over all extensions $\hat{\chi}$ of χ, so that

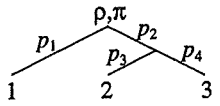
$$\mathbb{P}[\chi|T, \pi, p] = \sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\chi} \mathbb{P}[\hat{\chi}|T, \pi, p] . \tag{4}$$

Figure 5 shows a calculation of $\mathbb{P}[\chi|T, \pi, p]$ for a simple tree and character.

The origin of the upper bound of $(r-1)/r$ on p_e lies in the assumption that changes of colour along an edge take place under a continuous-time Markov process. This generates a mutation probability p_e—the probability of a net change across the edge—such that $0 < p_e < (r-1)/r$, depending on the "length" of the edge; for simplicity we include the endpoints so that our set of possible vectors is compact. The mutation probabilities thus give some indication of the relative lengths of time along each edge, with p_e tending to $(r-1)/r$ as the "length" of e increases.

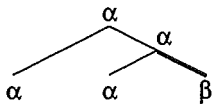
In a more general setting, we might allow each edge to be governed by an $r \times r$ *transition matrix* m^e whose $\alpha\beta$-entry gives the probability that a change from α to β occurs across e, given that the vertex nearer the root is coloured α. Under this model, (2) implies

$$\mathbb{P}[\hat{\chi}|T, \pi, \{m^e\}] = \pi(\hat{\chi}(\rho)) \prod_{e=(u,v)\in E(T)} m^e_{\hat{\chi}(u)\hat{\chi}(v)} . \tag{5}$$

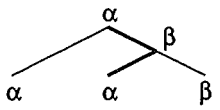


C={α,β}

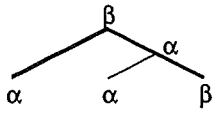
i	1	2	3
$\chi(i)$	α	α	β



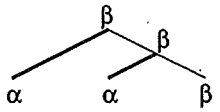
$$\mathbb{P}[\hat{\chi}|T,\pi,p] = \pi(\alpha)(1-p_1)(1-p_2)(1-p_3)p_4$$



$$\mathbb{P}[\hat{\chi}|T,\pi,p] = \pi(\alpha)(1-p_1)p_2p_3(1-p_4)$$



$$\mathbb{P}[\hat{\chi}|T,\pi,p] = \pi(\beta)p_1p_2(1-p_3)p_4$$



$$+ \quad \mathbb{P}[\hat{\chi}|T,\pi,p] = \pi(\beta)p_1(1-p_2)p_3(1-p_4)$$

$$= \mathbb{P}[\chi|T,\pi,p]$$

Figure 5: To calculate $\mathbb{P}[\chi|T,\pi,p]$ for the tree and character shown, sum $\mathbb{P}[\hat{\chi}|T,\pi,p]$ over all extensions $\hat{\chi}$ of χ. In the two colour case, each bi-coloured edge contributes a factor of p_e to $\mathbb{P}[\hat{\chi}|T,\pi,p]$; all other edges contribute a factor of $1-p_e$.

Note that in this case the product is taken over edges directed away from the root, as the matrices need not be symmetric. For our model the transition matrices are

$$m^e = \begin{pmatrix} 1-p_e & \frac{p_e}{r-1} & \cdots & \frac{p_e}{r-1} \\ \frac{p_e}{r-1} & 1-p_e & & \vdots \\ \vdots & & \ddots & \frac{p_e}{r-1} \\ \frac{p_e}{r-1} & \cdots & \frac{p_e}{r-1} & 1-p_e \end{pmatrix}. \tag{6}$$

Each diagonal entry is $1-p_e$ and all the off-diagonal entries are $p_e/(r-1)$.

2.3.1 Mutation probability along a path

Our first result is a generalisation to r colours of a result due to Hendy [9] in the two colour case.

Lemma 1 (Mutation probability along a path) *If u and v are vertices of T, then*

$$\mathbb{P}[\hat{\chi}(u) \neq \hat{\chi}(v)|T,\pi,p] = \frac{r-1}{r}(1-\prod_{e\in P}(1-\frac{r}{r-1}p_e)), \tag{7}$$

where P is the path between u and v.

Proof: The result may be proved by induction on the length of the path or using linear algebra. We give the linear algebra proof.

If u is an ancestor vertex of v then the transition matrix m^P for the path is the product of the matrices m^e along the path. Diagonalising the matrix in (6) we find

$$m^e = H\mathrm{diag}(1, 1-\frac{r}{r-1}p_e, \ldots, 1-\frac{r}{r-1}p_e)H^{-1}, \tag{8}$$

where

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & -1 & 1 & 1 & \cdots & 1 \\ 1 & 0 & -2 & 1 & \cdots & 1 \\ 1 & 0 & 0 & -3 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & 1 \\ 1 & 0 & 0 & \cdots & 0 & 1-r \end{pmatrix}. \tag{9}$$

Hence

$$m^P = H\mathrm{diag}(1, \prod_{e\in P}(1-\frac{r}{r-1}p_e), \ldots, \prod_{e\in P}(1-\frac{r}{r-1}p_e))H^{-1}, \tag{10}$$

which by (8) is a matrix of the form in (6) with mutation probability

$$p_P = \frac{r-1}{r}(1-\prod_{e\in P}(1-\frac{r}{r-1}p_e)). \tag{11}$$

Now if u and v are arbitrary vertices of T, let w be their most recent common ancestor, π_w the distribution of colours at w and p_1, p_2 the path mutation probabilities from w to u and v respectively (see figure 6). The probability that u and v are the same colour is

$$\begin{aligned} \mathbb{P}[\hat{\chi}(u) = \hat{\chi}(v)|T,\pi,p] &= \sum_{\alpha\in C}[\pi_w(\alpha)(1-p_1)(1-p_2) + \sum_{\substack{\gamma\in C \\ \gamma\neq\alpha}}\pi_w(\gamma)\frac{p_1p_2}{(r-1)^2}] \\ &= (1-p_1)(1-p_2) + \frac{p_1p_2}{r-1} \\ &= 1-p_1-p_2+\frac{r}{r-1}p_1p_2, \end{aligned}$$

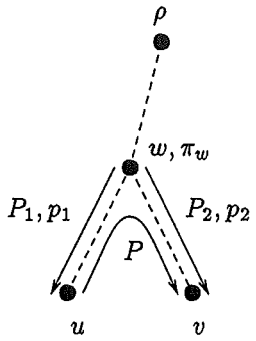


Figure 6: If u and v are not related by direct descent, break P as shown into the paths P_1 and P_2 with associated path mutation probabilities p_1, p_2, where w is the most recent common ancestor of u and v.

since each $\pi_w(\gamma)$ appears $r-1$ times in the sum over α and γ, and $\sum_{\alpha \in C} \pi_w(\alpha) = 1$. Therefore

$$\begin{aligned} \mathbb{P}[\hat{\chi}(u) \neq \hat{\chi}(v) | T, \pi, p] &= p_1 + p_2 - \frac{r}{r-1} p_1 p_2 \\ &= \frac{r-1}{r}(1 - (1 - \frac{r}{r-1} p_1)(1 - \frac{r}{r-1} p_2)) \quad (12) \end{aligned}$$

and substituting $p_i = \frac{r-1}{r}(1 - \prod_{e \in P_i}(1 - \frac{r}{r-1} p_e))$ from (11) for each i we obtain (7).

Remark: Note that if u and v are not related by direct descent then the transition matrix for the path P is not necessarily of the form in (6). However, in the case of an even distribution of colours at the root it is easily checked that we do obtain such a matrix.

2.3.2 Dealing with the root

The presence of the root and the distribution π introduce complications we would rather be without. Most tree reconstruction methods return unrooted trees, so that the position of the root is generally not known. The distribution at the root is typically another unknown. We consider here two ways of obtaining some measure of $\mathbb{P}[\chi|T]$ on an unrooted tree closely related to T.

Firstly and most simply we could assume an even distribution of colours at the root, that is $\pi(\alpha) = 1/r \; \forall \alpha \in C$, to get

$$\mathbb{P}[\chi | T, \pi, p] = \frac{1}{r} \sum_{\hat{\chi} : \hat{\chi}|_{[n]} = \chi} \prod_{\substack{e=\{u,v\}: \\ \hat{\chi}(u) = \hat{\chi}(v)}} (1 - p_e) \prod_{\substack{e=\{u,v\}: \\ \hat{\chi}(u) \neq \hat{\chi}(v)}} \frac{p_e}{r-1} \; . \quad (13)$$

This expression is independent of the location of the root so we may treat the tree as unrooted. Where the root has degree two this does not give us an unrooted phylogenetic tree since we do not allow non-root vertices to have degree two, but in this case we may use the remark following Lemma 1 to collapse the root and the two edges incident with it to a single edge with the path mutation probability

$$p_P = \frac{r-1}{r}(1-(1-\frac{r}{r-1}p_1)(1-\frac{r}{r-1}p_2)), \tag{14}$$

thereby obtaining an unrooted phylogenetic tree (see figure 7).

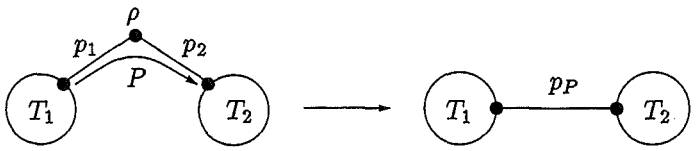


Figure 7: Deleting the root. The circles marked T_1, T_2 denote rooted subtrees.

A second approach is to allow a transitive subgroup G of S_C, the symmetric group on C, to act naturally on the set of characters on n leaves according to $(\sigma\chi)(i) = \sigma(\chi(i))\ \ \forall \sigma \in G, i \in [n]$. Summing over an orbit $G\chi$ we have

$$\begin{aligned}\mathbb{P}[G\chi|T,\pi,p] &= \sum_{\sigma\in G}\sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\sigma\chi} \mathbb{P}[\hat{\chi}|T,\pi,p] \\ &= \sum_{\sigma\in G}\sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\sigma\chi} \pi(\hat{\chi}(\rho)) \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)=\hat{\chi}(v)}} (1-p_e) \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)\neq\hat{\chi}(v)}} \frac{p_e}{r-1}\ .\end{aligned}$$

If $\hat{\chi}$ is an extension of χ then $\sigma\hat{\chi}$ is an extension of $\sigma\chi$, so we may exchange the order of summation to get

$$\begin{aligned}\mathbb{P}[G\chi|T,\pi,p] &= \sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\chi}\sum_{\sigma\in G} \pi(\sigma\hat{\chi}(\rho)) \prod_{\substack{e=\{u,v\}:\\ \sigma\hat{\chi}(u)=\sigma\hat{\chi}(v)}} (1-p_e) \prod_{\substack{e=\{u,v\}:\\ \sigma\hat{\chi}(u)\neq\sigma\hat{\chi}(v)}} \frac{p_e}{r-1} \\ &= \sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\chi} \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)=\hat{\chi}(v)}} (1-p_e) \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)\neq\hat{\chi}(v)}} \frac{p_e}{r-1} \sum_{\sigma\in G} \pi(\sigma\hat{\chi}(\rho))\ ,\end{aligned}$$

the last following from $\sigma\hat{\chi}(u) = \sigma\hat{\chi}(v)$ if and only if $\hat{\chi}(u) = \hat{\chi}(v)$. Finally, G is transitive and $\sum_{\alpha\in C}\pi(\alpha) = 1$ so that $\sum_{\sigma\in G}\pi(\sigma\hat{\chi}(\rho)) = |G|/r$, and

$$\mathbb{P}[G\chi|T,\pi,p] = \frac{|G|}{r}\sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\chi} \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)=\hat{\chi}(v)}} (1-p_e) \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)\neq\hat{\chi}(v)}} \frac{p_e}{r-1}\ . \tag{15}$$

Thus $\mathbb{P}[G\chi|T,\pi,p]$ is independent of π and we have an expression very similar to that in (13) where we considered the even distribution case. Where the root has degree two we may again collapse it and its incident edges as in figure 7 to obtain an unrooted tree with $\mathbb{P}[G\chi|T,\pi,p]$ unchanged.

Where $G = S_C$ we note that $G\chi$ is the set of all characters inducing the partition $\{\chi^{-1}(\{\alpha\}) : \alpha \in C\}$ of $[n]$.

Both of the above methods allow us to associate an unrooted tree with each rooted tree. Where the root has degree greater than two, the unrooted tree is identical to the rooted tree except that we no longer distinguish the root; where the root has degree two we collapse the two edges incident with it to a single edge. The mutation probability across this edge is the net mutation probability across the path formed by the two collapsed edges. The sum over extensions of χ is common to both (13) and (15), motivating the following definition:

Definitions 4 *For an r-colour character χ and an unrooted tree T we define*

$$\mathbb{P}'[\chi|T,p] = \sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\chi} \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)=\hat{\chi}(v)}} (1-p_e) \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)\neq\hat{\chi}(v)}} \frac{p_e}{r-1} \tag{16}$$

In the case of an even distribution of colours at the root $\mathbb{P}'[\chi,T,p]$ may be interpreted as the conditional probability of generating χ on the rooted tree from which T was obtained, given that leaf 1 (say) is coloured $\chi(1)$. This follows from (13) and the fact that an even distribution of colours at the root induces an even distribution of colours at each leaf.

For much of the remainder of this discussion we consider $\mathbb{P}'[\chi|T,p]$ and unrooted trees.

3 Bounding $\mathbb{P}'[\chi|T,p]$

Penny et. al. [11] have shown that in the two colour case,

$$\max_p\{\mathbb{P}'[\chi|T,p]\} = 2^{-\ell(\chi,T)}. \tag{17}$$

A major result of this project is an extension of this result to r colours:

Theorem 3 (Upper bound for $\mathbb{P}'[\chi|T,p]$) *If χ is an r-colour character and T is an unrooted tree, then*

$$\max_p\{\mathbb{P}'[\chi|T,p]\} = r^{-\ell(\chi,T)}. \tag{18}$$

As a corollary to the proof of Theorem 3 we have the following:

Theorem 4 *A minimal extension of an r-colour character χ on T is uniquely determined by χ and the set of edges it bi-colours. That is, if χ_1 and χ_2 are minimal extensions of χ and*

$$\{e = \{u, v\} \in E(T) : \chi_1(u) \neq \chi_1(v)\} = \{e = \{u, v\} \in E(T) : \chi_2(u) \neq \chi_2(v)\}, \tag{19}$$

then $\chi_1 = \chi_2$.

Before proving Theorem 3 we consider the proof in the two colour case. If P is a path in T joining leaves i and j, put

$$\phi(P) = \begin{cases} 1 & \text{if } \chi(i) \neq \chi(j) \\ 0 & \text{if } \chi(i) = \chi(j) \end{cases} . \tag{20}$$

By Menger's theorem there is a set $\{P_1, \ldots, P_\ell\}$ of ℓ edge-disjoint paths such that $\phi(P_i) = 1$ for $i = 1, \ldots, \ell$, and by Lemma 1

$$\mathbb{P}[\phi(P) = 1 | T, p] = p_P = \frac{1}{2}(1 - \prod_{e \in P_i} (1 - 2p_e)) \leq \frac{1}{2}. \tag{21}$$

Our first two assumptions on page 7 imply that changes of colour on different edges are independent, so that the $\phi(P_i)$ are independent variables since the P_i are edge disjoint. Hence

$$\begin{aligned} \mathbb{P}[\phi(P_1) = 1, \ldots, \phi(P_\ell) = 1 | T, p] &= \prod_{i=1}^{\ell} \mathbb{P}[\phi(P_i) = 1 | T, p] \\ &\leq 2^{-\ell} , \end{aligned}$$

implying $\mathbb{P}'[\chi | T, p] \leq 2^{-\ell}$. To complete the proof, a vector p such that $\mathbb{P}'[\chi | T, p] = 2^{-\ell}$ is exhibited.

The strong use of Menger's theorem, in the requirement that the paths be edge disjoint, effectively prevents any natural extension of this proof to r colours as an optimal Erdős-Székely path system may have intersecting paths.

3.1 Proof of Theorems 3 and 4

We begin by reducing to the case where p_e is either 0 or $(r-1)/r$ for every edge e of T. For notational convenience put

$$M(T) = \{p \in [0, (r-1)/r]^{|E(T)|} : p_e \in \{0, (r-1)/r\} \forall e \in E(T)\} \tag{22}$$

and define

$$E(p) = \{e \in E(T) : p_e = (r-1)/r\} \tag{23}$$

for each $p \in M(T)$, and

$$E(\hat{\chi}) = \{e = \{u, v\} \in E(T) : \hat{\chi}(u) \neq \hat{\chi}(v)\} \tag{24}$$

(the *change set* or *bi-coloured set* of $\hat{\chi}$) for each colouration $\hat{\chi}$ of T.

Let χ be an r-colour character of length ℓ on an unrooted tree T. For this section we consider χ and T to be fixed and write $\mathbf{P}(p)$ for $\mathbf{P}'[\chi|T,p]$, to emphasise the view of $\mathbf{P}'[\chi|T,p]$ as a function of p. Thus

$$\mathbf{P}(p) = \sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\chi} \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)=\hat{\chi}(v)}} (1-p_e) \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)\neq\hat{\chi}(v)}} \frac{p_e}{r-1} \ . \tag{25}$$

Note that for each edge e of T, p_e occurs in each term in the sum in (25) exactly once. Let $p \in [0,(r-1)/r]^{|E(T)|}$. Choosing $e' \in E(T)$ and fixing p_e for $e \in E(T)\setminus\{e'\}$ (so that we regard $\mathbf{P}(p)$ as a function of $p_{e'}$), we therefore obtain a polynomial of degree at most one in $p_{e'}$. On a closed interval, the extreme values of such a polynomial occur at the end points, so there is a vector p' of mutation probabilities such that

$$p'_e = \begin{cases} p_e & \text{if } e \neq e' \\ 0 \text{ or } \frac{r-1}{r} & \text{if } e = e' \end{cases} \tag{26}$$

and

$$\mathbf{P}(p) \leq \mathbf{P}(p'). \tag{27}$$

Carrying out this process for each edge of T in turn, we eventually arrive at a vector p'' such that $p'' \in M(T)$ and

$$\mathbf{P}(p) \leq \mathbf{P}(p''). \tag{28}$$

We have established the following lemma:

Lemma 2 $\max_p\{\mathbf{P}'[\chi|T,p]\}$ *is realised by some* $p \in M(T)$.

Now let $p \in M(T)$. Each extension $\hat{\chi}$ of χ contributes a term

$$\mathbf{P}'[\hat{\chi}|T,p] = \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)=\hat{\chi}(v)}} (1-p_e) \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)\neq\hat{\chi}(v)}} \frac{p_e}{r-1} \tag{29}$$

to $\mathbf{P}(p)$. If there is an edge $e=\{u,v\}$ for which $\hat{\chi}(u) \neq \hat{\chi}(v)$ and $p_e = 0$, then a factor of zero occurs in the right-hand product in (29) and we have $\mathbf{P}'[\hat{\chi}|T,p] = 0$. Hence we need only sum over extensions $\hat{\chi}$ such that $E(\hat{\chi}) \subseteq E(p)$. Further, if $E(\hat{\chi}) \subseteq E(p)$, then each edge $e=\{u,v\}$ contributes a factor

$$m^e_{\hat{\chi}(u)\hat{\chi}(v)} = \begin{cases} \frac{p_e}{r-1} = \frac{1}{r} & \text{if } \hat{\chi}(u) \neq \hat{\chi}(v) \\ 1-p_e = \frac{1}{r} & \text{if } \hat{\chi}(u) = \hat{\chi}(v) \text{ and } p_e = \frac{r-1}{r} \\ 1-p_e = 1 & \text{if } \hat{\chi}(u) = \hat{\chi}(v) \text{ and } p_e = 0 \end{cases} \tag{30}$$

to $\mathbf{P}'[\hat{\chi}|T,p]$. Thus each edge for which $p_e = (r-1)/r$ contributes a factor of $1/r$ to $\mathbf{P}'[\hat{\chi}|T,p]$, and all other edges a factor of 1, so we have:

Lemma 3 *If $p \in M(T)$, then*

$$\mathbb{P}(p) = \frac{1}{r^{|E(p)|}} |\{\hat{\chi} : \hat{\chi}|_{[n]} = \chi, E(\hat{\chi}) \subseteq E(p)\}|. \tag{31}$$

By Lemma 3, to calculate $\mathbb{P}(p)$ for $p \in M(T)$, we must count the number of extensions $\hat{\chi}$ of χ for which $E(\hat{\chi}) \subseteq E(p)$. With a view to proving Theorem 3, we would like to show that

$$|\{\hat{\chi} : \hat{\chi}|_{[n]} = \chi, E(\hat{\chi}) \subseteq E(p)\}| \leq r^{|E(p)|-\ell}, \tag{32}$$

with this bound attained by some $p \in M(T)$. Since it may be the case that there are no extensions with $E(\hat{\chi}) \subseteq E(p)$ (this will certainly be the case if $|E(p)| < \ell$), we make the following definition:

Definitions 5 (χ-viable) *$S \subseteq E(T)$ is* χ-viable *or* viable for χ on T *if there is an extension $\hat{\chi}$ of χ such that $E(\hat{\chi}) \subseteq S$.*

Let S be viable for χ on T, $\hat{\chi}$ such that $E(\hat{\chi}) \subseteq S$, and put $k = |S|$. Deleting S from T, which we denote by $T \setminus S$, will divide T into $k+1$ connected components, and $\hat{\chi}$ must be constant on each of these since $E(\hat{\chi}) \subseteq S$. In particular, if v is a vertex belonging to a component containing a leaf i of T, then we must have $\hat{\chi}(v) = \chi(i)$. However, on components that do not contain a leaf of T, $\hat{\chi}$ may take any of the r colours in C. Since $\hat{\chi}$ is completely determined by the colour of each connected component of $T \setminus S$, it follows that if there are λ components not containing a leaf of T then there are precisely r^λ extensions of χ such that $E(\hat{\chi}) \subseteq S$.

Definitions 6 (internal and external components) *If $S \subseteq E(T)$, a connected component of $T \setminus S$ that does not contain a leaf of T is an* internal component. *A connected component that does contain a leaf of T is an* external component.

Figure 8 shows an example of a tree and character with a set of χ-viable edges deleted. The deleted edges are shown as dashed lines and the connected components are circled. There is one internal component.

The inequality (32) follows from the above arguments and the following theorem:

Theorem 5 *Let χ be an r-colour character of length ℓ on T. If S is viable for χ on T and $|S| = k$, then $T \setminus S$ has at most $k - \ell$ internal components.*

Proof: Let $\mathcal{P} = \{P_1, \ldots, P_\ell\}$ be an optimal Erdős-Székely path system for χ on T. Since S is χ-viable and each path in P joins differently coloured leaves, there must be at least one edge of S on each P_i, so that following [5] we may

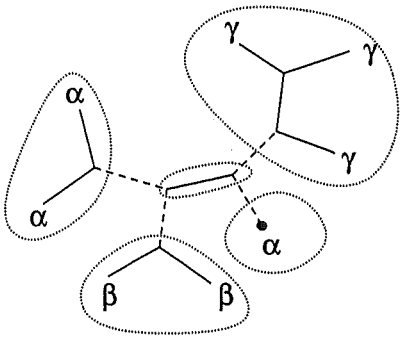


Figure 8: An example of $T \setminus S$.

define $f : \mathcal{P} \mapsto E(T)$ by $f(P) = e$ if e is the last edge in S on P. The conditions for an Erdős-Székely path system then imply that f is one-to-one. For if $f(P_i) = f(P_j) = e = (u, v)$, then both P_i and P_j must use e in the direction from u to v, and since no edges in S lie on that part of P_i from v to $t(P_i)$, nor on that part of P_j from v to $t(P_j)$, $t(P_i)$ and $t(P_j)$ belong to the same connected component of $T \setminus S$. This contradicts the fact that S is χ-viable and $\chi(t(P_i)) \neq \chi(t(P_j))$ for intersecting paths P_i and P_j of an Erdős-Székely path system.

Consider $f(\mathcal{P})$. We have $|f(\mathcal{P})| = \ell$ since f is one-to-one and $|P| = \ell$, so that $T \setminus f(\mathcal{P})$ has $\ell + 1$ connected components. We show that each component of $T \setminus f(\mathcal{P})$ contains a leaf of T. When the remaining edges in S are deleted, there will still be at least $\ell + 1$ connected components containing a leaf of T, so that $T \setminus S$ has at most $k + 1 - (\ell + 1) = k - \ell$ internal components.

Let $v \in V(T)$. The connected component of v in $T \setminus f(\mathcal{P})$ will contain a leaf of T if there is a walk from v to a leaf that does not cross an edge of $f(\mathcal{P})$. Let W be a walk from v to any leaf of T. If W does not cross any edges in $f(\mathcal{P})$ we are done; otherwise there is some $P_i \in \mathcal{P}$ such that $f(P_i) = e_i = (u_i, v_i)$ is the first edge in $f(\mathcal{P})$ that W crosses. We consider two cases, according to the direction in which W crosses e_i.

If W crosses e_i in the opposite direction to P_i (so that W arrives at v_i before u_i), then since e_i is the last edge in $f(\mathcal{P})$ on P_i, the path W' formed by following W as far as v_i and then traversing P_i forwards from v_i is a path from v to the leaf $t(P_i)$ that does not cross any edges in $f(\mathcal{P})$ (see figure 9).

Otherwise, if W crosses e_i in the same direction as P_i, let W' be the path formed by following W as far as u_i and then tracing P_i backwards towards $s(P_i)$. If there are no other edges in $f(\mathcal{P})$ on P_i then we obtain a path from v to a leaf that does not cross any edges in $f(\mathcal{P})$; otherwise there is $e_j = (u_j, v_j) = f(P_j)$ such that e_j is the first edge of $f(\mathcal{P})$ on W'. Since P_i and P_j must cross e_j in the same direction, W' crosses e_j in the opposite direction to P_j. We are now in the same position as in figure 9, so that the path W'' formed by following

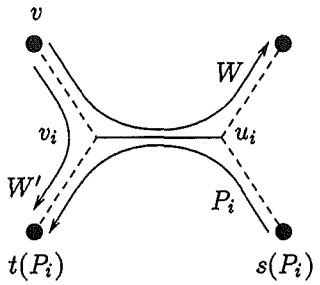


Figure 9: The walk W from v crosses $f(P_i)$ in the opposite direction to P_i. Trace P_i forwards to $t(P_i)$ to obtain W'.

W' as far as v_j and then tracing P_j forwards to $t(P_j)$ joins v to a leaf without crossing any edges in $f(\mathcal{P})$ (see figure 10).

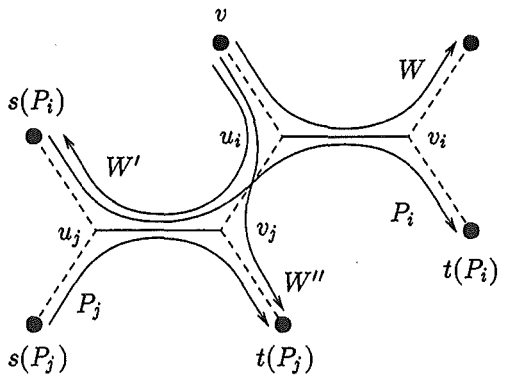


Figure 10: The walk W from v arrives at u_i before v_i. Retrace P_i towards $s(P_i)$ to form W'. If W' arrives at $s(P_i)$ without crossing any edges of $f(\mathcal{P})$ we are done; otherwise, we arrive at v_j and may trace P_j forwards to $t(P_j)$ without crossing any edges of $f(\mathcal{P})$, forming W''.

Therefore, given any vertex v of T there is a path joining it to a leaf of T that does not cross any edges in $f(\mathcal{P})$, and we conclude that each connected component of $T \setminus f(\mathcal{P})$ is external. The result follows.

Theorem 4 follows as an easy consequence of Theorem 5:

Corollary 1 (Theorem 4) *Let χ_1, χ_2 be minimal extensions of an r-colour character χ on T. Then $E(\chi_1) = E(\chi_2) \Leftrightarrow \chi_1 = \chi_2$.*

Proof: If χ_1 is a minimal extension of χ then $E(\chi_1)$ is a χ-viable set of cardinality $\ell(\chi, T)$. Then, by Theorem 5, $T \setminus E(\chi_1)$ has no internal components so that there are exactly $r^0 = 1$ extensions $\hat{\chi}$ of χ such that $E(\hat{\chi}) \subseteq E(\chi_1)$, namely $\hat{\chi} = \chi_1$. Hence if $E(\chi_1) = E(\chi_2)$ then $\chi_1 = \chi_2$.

Theorem 5 establishes the inequality (32), proving $\mathbb{P}'[\chi|T,p] \leq r^{-\ell(\chi,T)}$. To complete the proof of Theorem 3, we must exhibit a vector of probabilities p such that $\mathbb{P}(p) = r^{-\ell}$. The vector $p^{\hat{\chi}}$ defined by

$$p^{\hat{\chi}}_{\{u,v\}} = \begin{cases} \frac{r-1}{r} & \text{if } \hat{\chi}(u) \neq \hat{\chi}(v) \\ 0 & \text{if } \hat{\chi}(u) = \hat{\chi}(v) \end{cases} \tag{33}$$

is easily seen to be such a vector whenever $\hat{\chi}$ is a minimal extension of χ and we have our result.

3.2 Bounding $\mathbb{P}[\chi|T,\pi,p]$

We consider here briefly applying Theorem 3 to the problem of bounding $\mathbb{P}[\chi|T,\pi,p]$. For an arbitrary root distribution, we have

$$\begin{aligned} \mathbb{P}[\chi|T,\pi,p] &\leq \max_{\alpha \in C} \pi(\alpha)\, \mathbb{P}'[\chi|T,p] \\ &\leq \max_{\alpha \in C} \pi(\alpha)\, r^{-\ell(\chi,T)}. \end{aligned} \tag{34}$$

This bound will certainly be sharp if there is minimal extension $\hat{\chi}$ of χ such that $\pi(\hat{\chi}(\rho)) = \max_{\alpha \in C} \pi(\alpha)$. In the special case of an even distribution of colours at the root, we have

$$\mathbb{P}[\chi|T,\pi,p] \leq r^{-\ell(\chi,T)-1}, \tag{35}$$

with this bound achieved by vectors $p^{\hat{\chi}}$ for minimal extensions $\hat{\chi}$.

4 Realising the upper bound in the two colour case

Having found an upper bound for $\mathbb{P}'[\chi|T,p]$, it is natural to ask under what circumstances this bound is achieved. In this section we give a partial answer to this question, answering it in the case $r = 2$. If $\hat{\chi}$ is a minimal extension of χ then $\mathbb{P}'[\chi|T,p^{\hat{\chi}}] = r^{-\ell(\chi,T)}$, where $p^{\hat{\chi}}$ is as defined above, and for $r = 2$ this turns out to be a complete characterisation of the vectors p maximising $\mathbb{P}'[\chi|T,p]$:

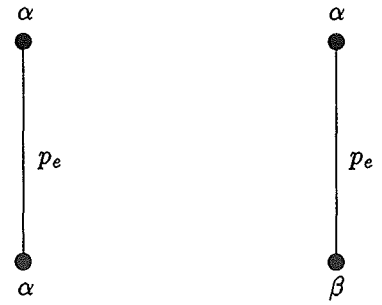
Theorem 6 *If χ is a binary character then p maximises $\mathbb{P}'[\chi|T,p]$ if and only if $p = p^{\hat{\chi}}$ for some minimal extension $\hat{\chi}$ of χ, where*

$$p^{\hat{\chi}}_{\{u,v\}} = \begin{cases} \frac{1}{2} & \text{if } \hat{\chi}(u) \neq \hat{\chi}(v) \\ 0 & \text{if } \hat{\chi}(u) = \hat{\chi}(v). \end{cases} \tag{36}$$

The backward direction of Theorem 6 is already established; we prove the forward direction in two stages, first establishing it for binary trees, and then reducing the general case to that where T is binary.

4.1 Proof of Theorem 6 for binary trees

The proof is by induction on n, the number of leaves of T. Consider $n = 2$, for which there are two possible characters χ up to permutation (see figure 11). Clearly $\mathbb{P}[\chi|T,p]$ is maximised in the first case only if $p_e = 0$, and in the second



$\mathbb{P}'[\chi|T,p] = 1 - p_e$ $\qquad$ $\mathbb{P}'[\chi|T,p] = p_e$

Figure 11: The two possible characters when $n = 2$.

only if $p_e = 1/2$.

Suppose the result is true for binary trees on $n-1$ leaves, where $n \geq 3$. Let T be a binary tree on n leaves, χ a character of length ℓ on T, and suppose that p is such that $\mathbb{P}'[\chi|T,p]$ is maximised. Since T is binary, it has a pair of adjacent pendant edges, that is a pair of edges $\{u,v\}$ and $\{u,v'\}$ such that v and v' are leaves of T (see figure 12). We consider two cases: $\chi(v) = \chi(v')$, and $\chi(v) \neq \chi(v')$.

Case 1: $\chi(v) = \chi(v')$.

Without loss of generality, $\chi(v) = \chi(v') = \alpha$. Let T' be the tree on $n-1$ leaves obtained by deleting $\{u,v\}$ and $\{u,v'\}$ from T, χ_α the character on the

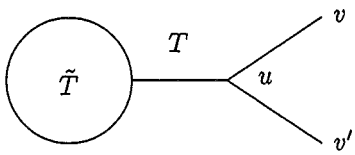


Figure 12: A tree T with a pair of adjacent pendant edges, $\{u, v\}$ and $\{u, v'\}$. The circle marked $\tilde{T}$ denotes a rooted subtree.

leaves of T' such that χ_α agrees with χ on their common leaves and $\chi_\alpha(u) = \alpha$, and define χ_β similarly. For convenience put $e = \{u, v\}$, $e' = \{u, v'\}$ and let the vertex w and edge e'' be as shown in figure 13.

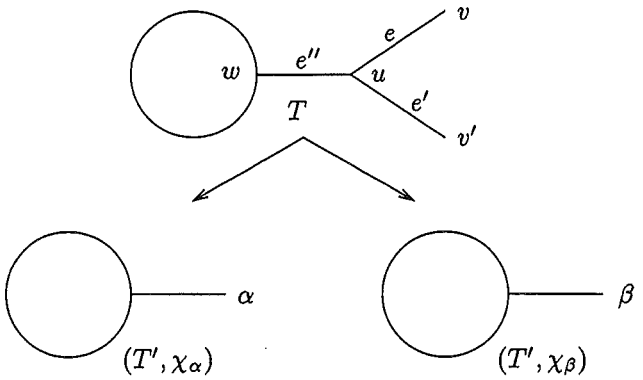


Figure 13: The trees T, T' and characters χ_α, χ_β.

Then

$$\mathbb{P}'[\chi|T, p] = (1 - p_e)(1 - p_{e'})\, \mathbb{P}'[\chi_\alpha|T', p] + p_e p_{e'}\, \mathbb{P}'[\chi_\beta|T', p]. \qquad (37)$$

Now if $\hat{\chi}$ is a minimal length extension of χ on T then $\hat{\chi}(u) = \alpha$; for if $\hat{\chi}(u) = \alpha$ we get no changes on e, e' and e'' if $\hat{\chi}(w) = \alpha$, and one change if $\hat{\chi}(w) = \beta$, while if $\hat{\chi}(u) = \beta$ we get two or three changes depending on whether $\hat{\chi}(w)$ equals α or β. It follows that χ_α has length ℓ on T'.

However χ_β may have length less than ℓ. For if $\bar{\chi}$ is an extension of χ such that $\bar{\chi}(u) = \beta$, then $\bar{\chi}$ is not a minimal length extension of χ and so has changing number at least $\ell + 1$. But two of these changes occur on e and e', which are deleted in forming T' and χ_β, so that $ch(\bar{\chi}|_{V(T')}) \geq \ell - 1$. Hence χ_β may have length less than ℓ but the decrease is by at most one. (For an example showing that this can in fact occur, see figure 14).

By Theorem 3, $\mathbb{P}'[\chi_\alpha|T', p] \leq 2^{-\ell}$ and $\mathbb{P}'[\chi_\beta|T', p] \leq 2^{-\ell+1}$ so that

$$\mathbb{P}'[\chi|T, p] \quad \leq \quad (1 - p_e)(1 - p_{e'})\, 2^{-\ell} + p_e p_{e'}\, 2^{-\ell+1}$$

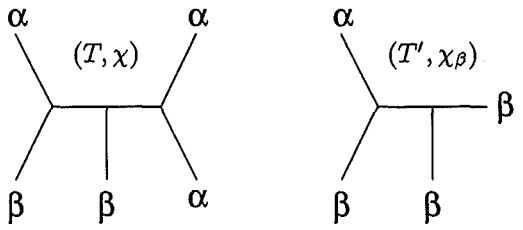


Figure 14: $\ell(\chi_\beta, T')$ may be less than $\ell(\chi, T)$. Here $\ell(\chi, T) = 2$ but $\ell(\chi_\beta, T') = 1$.

$$\begin{aligned} &= 2^{-\ell}((1-p_e)(1-p_{e'}) + 2p_e p_{e'}) \\ &= 2^{-\ell}(1 - p_e - p_{e'} + 3p_e p_{e'}). \end{aligned} \tag{38}$$

Consider $1 - p_e - p_{e'} + 3p_e p_{e'} = 1 - p_{e'} + p_e(3p_{e'} - 1)$. If $p_e = 0$ then

$$1 - p_e - p_{e'} + 3p_e p_{e'} = 1 - p_{e'} \leq 1,$$

with equality if and only if $p_{e'} = 0$. If $p_e > 0$ and $0 \leq p_{e'} < 1/3$ then $p_e(3p_{e'} - 1) < 0$ so $1 - p_e - p_{e'} + 3p_e p_{e'} < 1$. Finally, if $1/3 \leq p_{e'} \leq 1/2$ then $1 - p_{e'} \leq 2/3$ and $p_e(3p_{e'} - 1) \leq 1/4$ so that

$$1 - p_e - p_{e'} + 3p_e p_{e'} \leq \frac{2}{3} + \frac{1}{4} = \frac{11}{12} < 1. \tag{39}$$

Hence $1 - p_e - p_{e'} + 3p_e p_{e'} \leq 1$ with equality if and only if $p_e = p_{e'} = 0$.

Since $\max_p\{\mathbb{P}'[\chi|T, p]\} = 2^{-\ell}$ and p maximises $\mathbb{P}'[\chi|T, p]$, we must have $p_e = p_{e'} = 0$. By the induction hypothesis, $\mathbb{P}'[\chi_\alpha|T', p] = 2^{-\ell}$ if and only if $p = p^{\hat{\chi}_\alpha}$ on T' for a minimal extension $\hat{\chi}_\alpha$ of χ_α. A minimal extension of χ_α extends naturally to a minimal extension of χ and $p_e = p_{e'} = 0$ so that $p = p^{\hat{\chi}}$ for a minimal extension $\hat{\chi}$ of χ on T.

Case 2: $\chi(v) \neq \chi(v')$.

Without loss of generality $\chi(v) = \alpha$ and $\chi(v') = \beta$. Let T', χ_α and χ_β again be as in figure 13. If $\hat{\chi}$ is a minimal extension of χ then $\hat{\chi}$ involves a change on exactly one of e, e' regardless of the colour assigned to u, so that $\ell(\chi_\alpha, T'), \ell(\chi_\beta, T') \geq \ell - 1$. Hence

$$\begin{aligned} \mathbb{P}'[\chi|T, p] &= (1-p_e)p_{e'}\,\mathbb{P}'[\chi_\alpha|T', p] + p_e(1-p_{e'})\,\mathbb{P}'[\chi_\beta|T', p] \\ &\leq 2^{-\ell+1}((1-p_e)p_{e'} + p_e(1-p_{e'})) \\ &= 2^{-\ell}(1 - (1-2p_e)(1-2p_{e'})). \end{aligned} \tag{40}$$

Since $1 - (1-2p_e)(1-2p_{e'}) \leq 1$ with equality if and only if at least one of $p_e, p_{e'} = 1/2$, either

(i) $p_e = 0$, $p_{e'} = 1/2$ and $\mathbb{P}'[\chi_\alpha|T', p] = 2^{-\ell+1}$;

(ii) $p_e = 1/2$, $p_{e'} = 0$ and $\mathbb{P}'[\chi_\beta|T', p] = 2^{-\ell+1}$;

or if $p_e p_{e'} \neq 0$ then

(iii) $\mathbb{P}'[\chi_\alpha|T', p] = \mathbb{P}'[\chi_\beta|T', p] = 2^{-\ell+1}$ and at least one of $p_e, p_{e'} = 1/2$.

Under the induction hypothesis (i) and (ii) have $p = p^{\hat{\chi}}$ for a minimal extension $\hat{\chi}$, so it remains to show that (iii) cannot occur. By the induction hypothesis, $\mathbb{P}'[\chi_\alpha|T', p] = \mathbb{P}'[\chi_\beta|T', p] = 2^{-\ell+1}$ occurs if and only if $E(p) = E(\hat{\chi}_\alpha) = E(\hat{\chi}_\beta)$ for minimal extensions $\hat{\chi}_\alpha$ and $\hat{\chi}_\beta$ of χ_α and χ_β respectively. Let i be a leaf of T' other than u, and without loss of generality assume $\chi(i) = \alpha$. Consider the number of changes that occur on the path P from i to u. Since $\hat{\chi}_\alpha(i) = \hat{\chi}_\alpha(u)$ an even number of changes must take place on this path under χ_α; but $\hat{\chi}_\beta(i) \neq \hat{\chi}_\beta(u)$ so that an odd number of changes must take place under χ_β. Hence $E(\hat{\chi}_\alpha) = E(\hat{\chi}_\beta)$ is not possible, so that (iii) cannot occur and the theorem is proved for binary trees.

4.2 Splits and refinement

In this section we introduce some concepts required for an auxiliary theorem (Theorem 9) that will allow us to reduce the general case to the case just proved.

Definitions 7 (Splits) *A* split *is a bi-partition of* $[n]$.

Splits arise naturally from trees and are a convenient method of comparing and dealing with trees. Given an edge e of T, we obtain a split corresponding to e by deleting e and grouping the leaves in each of the rooted subtrees thereby created (see figure 15). Doing this for each edge of T we obtain the set of splits of T, $\sigma(T)$. The following theorem (Buneman, [2]) shows that $\sigma(T)$ contains all the information contained in T:

Theorem 7 *A set Σ of splits is $\sigma(T)$ for a phylogenetic tree T if and only if*

1. $\{\{i\}, [n] \setminus \{i\}\} \in \Sigma$, $i = 1, \ldots, n$;
2. *For each pair $\{A, B\}, \{C, D\} \in \Sigma$, at least one of $A \cap C$, $A \cap D$, $B \cap C$ and $B \cap D$ is empty.*

Furthermore, $\sigma(T) = \sigma(T')$ if and only if $T = T'$.

A set of splits satisfying condition (2) above is said to be *pairwise compatible.*

Splits may be used to define a partial order on the set of trees on n leaves:

Theorem 8 *The order $\leq$ defined by*

$$T_1 \leq T_2 \iff \sigma(T_1) \subseteq \sigma(T_2) \tag{41}$$

is a partial order on the set of phylogenetic trees on n leaves. The maximal elements are the binary phylogenetic trees.

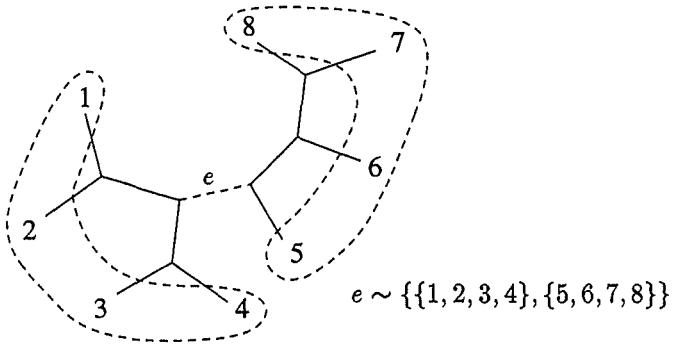


Figure 15: An edge and its corresponding split. Deleting the edge e from T, we obtain the split $\{\{1,2,3,4\},\{5,6,7,8\}\}$.

Definitions 8 (Refinement) *If $T_1 \leq T_2$ then T_2 is said to be a* refinement *of T_1.*

We may now state and prove the theorem required for the reduction of Theorem 6 to the binary tree case.

Theorem 9 *Let T be a tree and χ a binary character. There is a binary tree T' refining T such that $\ell(\chi,T') = \ell(\chi,T)$. T' may be chosen in such a way that the minimal extensions of χ on T' are in a natural bijective correspondence with the minimal extensions of χ on T.*

Proof: Let $\mathcal{P}$ be a set of $\ell = \ell(\chi,T)$ edge-disjoint paths joining differently coloured leaves, the existence of which is guaranteed by Menger's theorem. Form the sequence $T = T_1 < T_2 < \cdots$ refining T inductively as follows. Given T_i, choose $v \in V(T_i)$ of degree greater than or equal to four. If there is a path $P \in \mathcal{P}$ passing through v, choose e_1 and e_2 incident with v and lying on P; otherwise choose e_1 and e_2 incident with v arbitrarily. If $e_1 \sim \{A,B\}$, $e_2 \sim \{C,D\}$ and $A \cap C = \emptyset$, it is easily checked that $\{A \cup C, B \cap D\}$ is a split and that $\Sigma = \sigma(T_i) \cup \{\{A \cup C, B \cap D\}\}$ is pairwise compatible, so we may put $\sigma(T_{i+1}) = \Sigma$. Then $T_i < T_{i+1}$, and no path in $\mathcal{P}$ lies on the edge corresponding to $\{A \cup C, B \cap D\}$ so that $\mathcal{P}$ remains edge disjoint in T_{i+1} (see figure 16).

The new edge in T_{i+1} splits v into two vertices, one of degree three and one of degree one less than that of v, so this process must eventually terminate in a binary tree $T_m = T'$. $\mathcal{P}$ remains edge-disjoint in T' so by Menger's theorem we have $\ell(\chi,T') \geq \ell$. If $\hat{\chi}$ is a minimal extension of χ on T then we may obtain an extension $\bar{\chi}$ of χ on T' by identifying each vertex of T' with the vertex of T it was created from during the refinement process and requiring $\hat{\chi}$ and $\bar{\chi}$ to agree under this identification. An edge of T' is bi-coloured if and only if it

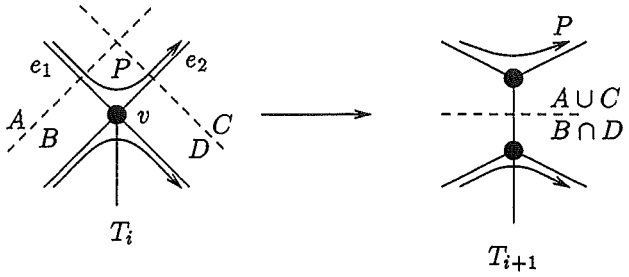


Figure 16: The refinement process. The splits are shown by dotted lines. None of the paths in $\mathcal{P}$ lie on the new edge of T_{i+1}, so that $\mathcal{P}$ remains edge disjoint in T_{i+1}. v splits into two vertices, one of degree three and the other of degree one less than the degree of v.

corresponds to a split of T and the corresponding edge of T is bi-coloured, so that $ch(\bar{\chi}) = ch(\hat{\chi}) = \ell$, implying $\ell(\chi, T') \leq \ell$ and hence equality.

Furthermore, every minimal extension of χ on T' arises in this way. Let $\bar{\chi}$ be such an extension. Since each path in $\mathcal{P}$ joins differently coloured leaves, there must be at least one change on each path. Moreover, $\mathcal{P}$ has cardinality $\ell(\chi, T')$, so there is exactly one change on each path and no changes on edges not on paths. Since none of the newly created edges lie on any of the paths, $\bar{\chi}$ must be constant on the set of vertices identified with a given vertex v of T, and we obtain a minimal extension $\hat{\chi}$ of χ on T by putting $\hat{\chi}(v)$ equal to this common colour.

4.3 The general case of Theorem 6

We now complete the proof of Theorem 6 in the general case.

Let T be a phylogenetic tree, χ a binary character and suppose p maximises $\mathbb{P}'[\chi|T, p]$. Let T' be a binary tree refining T as constructed in Theorem 9, and put $p'_{e'} = p_e$ if e and e' correspond to the same split σ of T, and $p'_{e'} = 0$ if e' does not correspond to a split of T. Then

$$\mathbb{P}'[\chi|T', p'] = \sum_{\hat{\chi}:\hat{\chi}|_{[n]}=\chi} \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)=\hat{\chi}(v)}} (1 - p'_e) \prod_{\substack{e=\{u,v\}:\\ \hat{\chi}(u)\neq\hat{\chi}(v)}} p'_e. \tag{42}$$

On newly created edges of T', $p'_e = 0$ so we need only sum over extensions for which no changes occur on newly created edges. Such an extension corresponds to an extension of χ on T, and it follows that

$$\mathbb{P}'[\chi|T', p'] = \mathbb{P}'[\chi|T, p] = 2^{-\ell(\chi,T)} = 2^{-\ell(\chi,T')}. \tag{43}$$

By the result proved for the binary tree case, $p' = p^{\bar{\chi}}$ for a minimal extension $\bar{\chi}$

of χ on T', and it follows from the construction of T' that $p = p^{\hat{\chi}}$ for a minimal extension $\hat{\chi}$ of χ on T.

4.4 Counter-examples for $r \geq 3$

Theorem 6 is not true when $r \geq 3$ even for binary trees, as the following counter-examples show. In part, this appears to be because r may be greater than or equal to the maximum degree of the internal vertices of T, making it easy to create an internal component from $E(\hat{\chi})$, $\hat{\chi}$ a minimal extension of χ, by deleting a single additional edge. Since phylogenetic trees are assumed to have no vertices of degree two, this does not occur for binary characters. However, if this requirement is dropped then Theorem 6 no longer holds. In particular, Theorem 6 does not carry over to rooted trees without some modification, as we allow the root to have degree two. We consider this case in section 4.5.

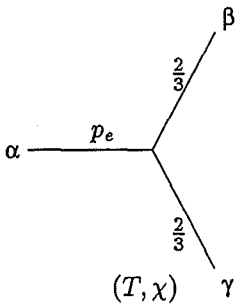


Figure 17: A counter-example to Theorem 6 for $r = 3$.

Example 1: A counter-example to Theorem 6 for $r = 3$ is illustrated by the star shaped tree in figure 17. We have $\ell(\chi, T) = 2$, since $ch(\hat{\chi}) = 2$ for all three possible extensions of χ on T. With p as shown, we have:

$$\begin{aligned} \mathbb{P}'[\chi|T,p] &= (1-p_e)\frac{1}{3}\frac{1}{3} + \frac{p_e}{2}(1-\frac{2}{3})\frac{1}{3} + \frac{p_e}{2}\frac{1}{3}(1-\frac{2}{3}) \\ &= \frac{1}{9} - \frac{1}{9}p_e + \frac{1}{18}p_e + \frac{1}{18}p_e \\ &= \frac{1}{9} = 3^{-2}, \end{aligned} \tag{44}$$

so that $\mathbb{P}'[\chi|T,p] = 3^{-\ell(\chi,T)}$ regardless of the value of p_e.

This example generalises readily to a counter-example for any $r \geq 3$ by considering the star-shaped tree on r leaves. This is the tree with vertices $\{0, 1, \ldots, r\}$ and edges $\{\{0,1\}, \{0,2\}, \ldots, \{0,r\}\}$.

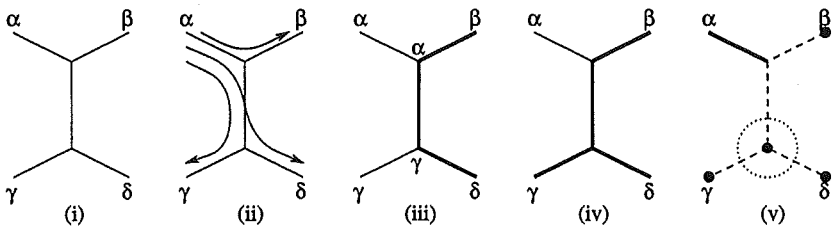


Figure 18: A counter-example to Theorem 6 for $r = 4$.

Example 2: A counter-example to Theorem 6 for $r = 4$ on a binary tree is illustrated in figure 18. Referring to this figure we have: (i) The tree T and character χ to be considered. (ii) An Erdős-Székely path system for χ on T. (iii) A colouration $\hat{\chi}$ of T of changing number 3. Bi-coloured edges are bolded. The path system in (ii) and this colouration, together with Theorem 2 imply $\ell(\chi, T) = 3$. (iv) A set S of edges (bolded). S is χ-viable since $E(\hat{\chi}) \subseteq S$. (iv) $T \setminus S$. Connected components are bolded and edges in S are dotted. There is one internal component (circled), so by Lemma 3 and the arguments following it, if $p \in M(T)$ with $E(p) = S$ then $\mathbb{P}'[\chi|T, p] = 4^{-4}.4^1 = 4^{-3} = 4^{-\ell(\chi,T)}$.

4.5 Theorem 6 for rooted trees

In this section we consider the form Theorem 6 takes for rooted trees. We do this only for the even distribution at the root case, as the only case for which the upper bound is sharp.

We have $\mathbb{P}[\chi|T, \pi, p] = \frac{1}{r}\mathbb{P}'[\chi|T', p']$, where T' and p' are the unrooted tree and mutation probability vector associated with T and p in section 2.3.2, so $\mathbb{P}[\chi|T, \pi, p]$ is maximised if and only if $\mathbb{P}'[\chi|T', p']$ is. Where the root has degree greater than two, $T = T'$and $p = p'$ so Theorem 6 holds as it stands. If the root has degree two, however, T' and p' are obtained by collapsing the edges incident with the root to a single edge with the path mutation probability

$$p_P = \frac{1}{2}(1 - (1 - 2p_1)(1 - 2p_2)), \tag{45}$$

as in figure 7, section 2.3.2. Vectors p maximising $\mathbb{P}[\chi|T, \pi, p]$ are obtained by re-inserting the root as in figure 19 and choosing p_1, p_2 such that (45) holds.

If $p_P = 0$ (so $p' = p^{\hat{\chi}}$ for a minimal extension $\hat{\chi}$ of χ on T' with $\hat{\chi}(u) = \hat{\chi}(v) = \alpha$ for some $\alpha \in C$) then we must choose $p_1 = p_2 = 0$. $\hat{\chi}$ may be extended to a minimal extension of χ on T by defining $\hat{\chi}(\rho) = \alpha$, and we then have $p = p^{\hat{\chi}}$.

If $p_P = 1/2$ (so $p' = p^{\hat{\chi}}$ for a minimal extension $\hat{\chi}$ of χ on T' with $\hat{\chi}(u) = \alpha$, $\hat{\chi}(v) = \beta$ for some $\alpha, \beta \in C$, $\alpha \neq \beta$). In this case we require only that at least one of p_1, p_2 equals $1/2$ in order to obtain $p_P = 1/2$. $\hat{\chi}$ may be extended to a

minimal extension of χ on T by putting either $\hat{\chi}(\rho) = \alpha$ or $\hat{\chi}(\rho) = \beta$, so in this case we have $p = p^{\hat{\chi}}$ except on one of the edges $\{u, \rho\}$ and $\{v, \rho\}$, on which p may take any value in the allowed range.

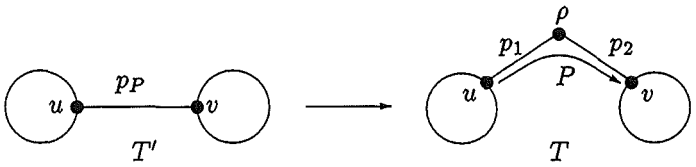


Figure 19: Re-inserting the root.

5 Applications to phylogenetic analysis

5.1 Equivalence of maximum parsimony and maximum likelihood with no common mechanism

Theorem 3 may be used to demonstrate the equivalence of two methods of phylogenetic inference, maximum parsimony and maximum likelihood with no common mechanism. Penny et. al. [11] state this result for the $r = 2$ case, and a simplified version of their result also appears in Goldman [8]. For a discussion of various methods of phylogenetic inference, see [8].

We consider only the model where the distribution of colours at the root is assumed to be uniform.

Maximum parsimony inference

Given a set $X = \{\chi_i\}$ of k r-colour characters, choose the unrooted tree or trees T (the "maximum parsimony tree(s)") minimising

$$\ell(X, T) = \sum_{i=1}^{k} \ell(\chi_i, T). \tag{46}$$

Interpreting $\ell(\chi_i, T)$ as the minimum number of mutations required for χ_i to evolve on T, $\ell(X, T)$ is the minimum total number of mutations required for the χ_i to evolve on T. Thus maximum parsimony chooses the trees on which the χ_i may evolve with as few mutations as possible overall.

Maximum likelihood inference

Definitions 9 (likelihood, Edwards [4]) *The* likelihood, $\mathbb{L}[H|R]$, *of the hypothesis H given data R and a specific model, is proportional to $\mathbb{P}[R|H]$, the constant of proportionality being arbitrary.*

A maximum likelihood method of inference chooses the hypothesis H maximising the likelihood function for the data R. For the model under consideration here, we may take the hypothesis to be the tree and mutation probability vector pair (T, p). The maximum likelihood method is then:

Given a set $X = \{\chi_i\}$ of k r-colour characters, choose the unrooted tree and vector pair or pairs (T, p) maximising

$$\mathbb{L}[(T,p)|X] = \mathbb{P}'[X|T,p] = \prod_{i=1}^{k} \mathbb{P}'[\chi_i|T,p]. \tag{47}$$

This is maximum likelihood with a common mechanism, since we require the same vector p to be used for each character; the tree estimated is the "maximum likelihood tree(s)". If we allow a different vector p for each character (so that the hypothesis becomes $(T, \{p_i\})$) we obtain maximum likelihood with no common mechanism:

Given a set $X = \{\chi_i\}$ of k r-colour characters, choose the unrooted tree and vector set pair or pairs $(T, \{p_i\})$ maximising

$$\mathbb{L}[(T,\{p_i\})|X] = \mathbb{P}'[X|T,\{p_i\}] = \prod_{i=1}^{k} \mathbb{P}'[\chi_i|T,p_i]. \tag{48}$$

For the model considered here where we assume an even distribution of colours at the root, we have the following result:

Theorem 10 *Maximum parsimony and maximum likelihood with no common mechanism are equivalent, in the sense that both choose the same tree or trees.*

Proof: The proof is the same as for the $r = 2$ case since it follows directly from Theorem 3. On any given tree T we have $\max_p \mathbb{P}'[\chi_i|T,p] = r^{-\ell(\chi_i,T)}$ so that

$$\max_{\{p_i\}} \mathbb{L}[(T,\{p_i\})|X] = \prod_{i=1}^{k} r^{-\ell(\chi_i,T)} = r^{-\sum_{i=1}^{k}\ell(\chi_i,T)} = r^{-\ell(X,T)}, \tag{49}$$

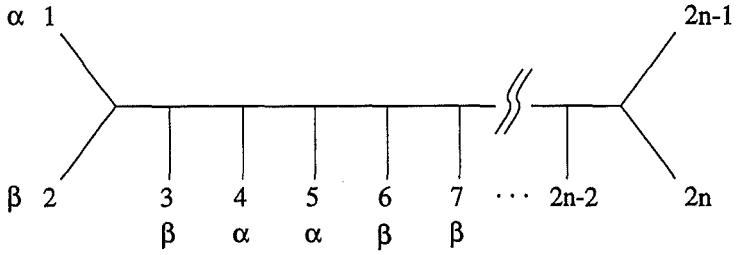
and therefore the maximum likelihood trees are precisely the maximum parsimony trees.

5.2 The maximum likelihood point is not unique

Maximum likelihood algorithms using a hill climbing method to maximise over the edge parameters on a given tree are effective in locating a stationary point of the likelihood function. The question then arises as to whether the stationary point found is a global or only a local maximum. Fukami and Tateno [7] claimed to have answered this question by showing that the likelihood function has a

unique stationary point. Steel [13] gave a simple counter-example to this claim, using a tree on four leaves for which the likelihood function had two extrema at widely separated points. The results of this project show that the likelihood function has more than one stationary point whenever the character considered has more than one minimal extension.

We have seen that $p^{\hat{\chi}}$ as defined in equation (33) maximises $\mathbb{P}'[\chi|T,p]$ whenever $\hat{\chi}$ is a minimal extension of χ. By Theorem 4, these vectors are distinct, so that there are at least as many vectors maximising $\mathbb{P}'[\chi|T,p]$ as there are minimal extensions of χ on T (by Theorem 6, exactly as many when $r = 2$). Since characters may have more than one minimal extension on a given tree (Steel [12] constructs a tree and character pair with a Fibonacci number of minimal extensions, see figure 20), it would appear that the likelihood point will in general not be unique.



$$\chi(i) = \begin{cases} \alpha & \text{if } i \equiv 0,1 \pmod 4, \\ \beta & \text{otherwise.} \end{cases}$$

Figure 20: The tree and character pair shown has a number of minimal extensions equal to the nth Fibonacci number.

6 Discussion

The main results of this project are the generalisation of the upper bound on $\mathbb{P}'[\chi|T,p]$ from two to r colours (Theorem 3) and the complete characterisation of the vectors maximising $\mathbb{P}'[\chi|T,p]$ when $r = 2$ (Theorem 6). These results answer the two main questions that formed the starting point of this investigation. Of additional interest are the characterisation of minimal extensions in terms of their bi-coloured sets (Theorem 4) and the existence of a binary tree refining a given tree on which a given character has the same length (Theorem 9). This latter result in particular may have applications outside the immediate sphere of interest.

Further work on this model could address the form Theorem 6 should take when $r \geq 3$ and examine in greater detail the effect of the distribution of colours at the root.

7 Acknowledgements

I would like to thank Dr Steel for his supervision of this project.

References

[1] J. A. Bondy and U. S. R. Murty. *Graph Theory with Applications.* Macmillan Press, London, 1976.

[2] P. Buneman. The recovery of trees from measures of dissimilarity. In F. R. Hodson, D. G. Kendall, and P. Tautu, editors, *Mathematics in the archaeological and historical sciences*, pages 387–395. Edinburgh University Press, 1971.

[3] J. A. Cavender. Taxonomy with confidence. *Mathematical Biosciences*, 40:270–280, 1978.

[4] A. W. F. Edwards. *Likelihood.* Cambridge University Press, Cambridge, 1972.

[5] P. L. Erdős and L. A. Székely. On weighted multiway cuts in trees. *Mathematical Programming*, 65:93–105, 1994.

[6] J. S. Farris. A probability model for inferring evolutionary trees. *Systematic Zoology*, 22:250–256, 1973.

[7] K. Fukami and Y. Tateno. On the maximum likelihood method for estimating molecular trees: Uniqueness of the likelihood point. *J. Mol. Evol.*, 28:460–464, 1989.

[8] N. Goldman. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Systematic Zoology*, 39(4):345–361, 1990.

[9] M. D. Hendy. A combinatorial description of the closest tree algorithm for finding evolutionary trees. *Discrete Mathematics*, 96:51–58, 1991.

[10] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian protein metabolism*, pages 21–132. Academic Press, New York, 1969.

[11] D. Penny, P. J. Lockhart, M. A. Steel, and M. D. Hendy. The role of models in reconstructing evolutionary trees. In R. W. Scotland, D. J. Siebert, and D. M. Williams, editors, *Models in Phylogeny Reconstruction,* Systematics Association Special Volume 52, pages 211–230. Clarendon Press, Oxford, 1994.

[12] M. A. Steel. Decompositions of leaf-colored binary trees. *Advances in Applied Mathematics*, 14:1–24, 1993.

[13] M. A. Steel. The maximum likelihood point for a phylogenetic tree is not unique. *Systematic Biology*, 43(4):560–564, 1994.